PRIMARY STRUCTURE OF THE CARBOHYDRATE-CONTAINING REGIONS OF THE CARBOXYL PROPEPTIDES OF TYPE I PROCOLLAGEN

D. M. PESCIOTTA, L. A. DICKSON+, A. M. SHOWALTER, E. F. EIKENBERRY, B. DE CROMBRUGGHE*, P. P. FIETZEK and B. R. OLSEN

Department of Biochemistry, CMDNJ-Rutgers Medical School, Piscataway, NJ 08854, +Department of Biochemistry, CMDNJ–New Jersey School of Osteopathic Medicine, Piscataway, NJ 08854 and **Laboratory of Molecular Biology, Division of Cancer Biology and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA*

Received 21 January 1981

1. Introduction

Procollagen, the biosynthetic precursor of collagen, contains asparagine-linked oligosaccharide units in the carboxyl propeptides of the proα-chains [1–6]. These oligosaccharide units contain *N*-acetylglucosamine and mannose. Based on molecular weight estimates of pronase-derived glycopeptides obtained from [^{3}H]-mannose-labeled type I procollagen, it has been concluded that the carboxyl propeptides of type I procollagen contain 8–9 monosaccharide moieties [6]. On the basis of direct determination of mannose and *N*-acetylglucosamine in carboxyl propeptides isolated and purified from type I procollagen, we had found that the proα1(I) propeptide contains 2 residues of *N*-acetylglucosamine and 9 residues of mannose and that the proα2 propeptide contains 2 residues of *N*-acetylglucosamine and 13 residues of mannose [3]. Based on a combination of cDNA and peptide sequence analysis, we were able to identify the attachment site for the oligosaccharide within the proα1(I) chain [7]. Here, we report on the amino acid sequence of the oligosaccharide attachment site within the proα2 chain. In addition, we have isolated and characterized the mannose-rich tryptic glycopeptides obtained from both the proα1(I) and proα2 carboxyl propeptides.

2. Material and methods

The carboxyl propeptide of chick type I procollagen, labeled with a mixture of ^{14}C-labeled amino acids was prepared and purified as in [3]. The purified peptide (7 mg in a typical experiment) was reduced and alkylated with sodium iodoacetate [3] and desalted on a Bio-Gel P2 (Bio-Rad Labs.) column [3]. Separation of the subunits of the carboxyl propeptide (C1,C2) was accomplished using a 1.5 × 10 cm column of CM-cellulose (CM-52; Whatman). The column was equilibrated with a 50 mM sodium acetate buffer (pH 4.2) containing 6 M urea and eluted with a 600 ml linear gradient of 0–0.2 M NaCl in the same buffer. The flow rate was 86 ml/h and 5 ml fractions were collected. The appropriate fractions were pooled, and the separated subunits were desalted on a Bio-Gel P2 column (Bio-Rad Labs.) equilibrated with 0.2 M ammonium bicarbonate and lyophilized [3].

Each propeptide (C1,C2) was dissolved in 0.2 M ammonium bicarbonate and digested with TosPhe Ch_2Cl-treated trypsin for 4.5 h at 37°C at a substrate/enzyme ratio of 20/1 (w/w). The reaction was stopped by adding a few drops of 0.5 N acetic acid and the samples were lyophilized.

The tryptic digests of the C1 and C2 propeptides were fractionated on a 0.5 × 1.7 cm column of con A–Sepharose (Pharmacia). The column was equilibrated with a 25 mM Tris–HCl buffer (pH 7.5) containing 0.1 M NaCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$, 0.1% Triton X-100, 0.2% SDS and 0.01% sodium azide [6]. The glycopeptides were eluted from the column with 0.05 M α-methyl-mannoside. The flow rate was 2.4 ml/h and 0.5 ml fractions were collected. The appropriate fractions were pooled and desalted on a 1.5 × 130 cm polyacrylamide column (Bio-Gel P6,

Address correspondence to B. R. O.

Bio-Rad Labs.) equilibrated with 0.2 M ammonium bicarbonate.

Neutral sugars were analyzed by gas chromatography as in [3] using a column packed with 3% SP-2340 on 100/120 Supelcoport in a Hewlett-Packard Model 5830A gas–liquid chromatograph. After hydrolysis of the peptides with 1 N HCl at 100°C for 9 h, 0.16 μmol arabinose were added as an internal standard.

Amino acid analyses were performed using a Beckman Model 121 MB amino acid analyzer. Peptides were sequenced by automated Edman degradation as in [11]. The PTH-amino acid derivatives were identified by high-pressure liquid chromatography and by amino acid analysis following back-hydrolysis. The eluate from the high pressure column was collected in 0.5 ml fractions and assayed for radioactivity by liquid scintillation counting.

The preparation of a hybrid bacterial plasmid containing a collagen proα2 cDNA sequence has been described [8]. The cDNA insert was isolated from the recombinant plasmid, pCOL 1, by digestion with *Bam*HI and *Hin*dIII followed by electrophoresis on a 5.2% polyacrylamide slab gel. The DNA was located by staining with ethidium bromide and recovered by electrophoretic elution.

The genomic clone, λgCOL 204, which contains a 16×10^3 basepair chicken DNA insert with sequences coding for the carboxy-terminal part of the proα2 chain was isolated from a genomic library using the cDNA insert of pCOL 1 as hybridization probe [9]. From a partial *Hin*dIII digest of λgCOL 204, a fragment of 10.5×10^3 basepairs containing a 3.8×10^3 and a 6.7×10^3 basepairs *Hin*dIII fragment was subcloned in the plasmid vector pBR322. This subclone, pgCOL 20, contains all the coding sequences for the carboxyl propeptide of proα2 chains.

The recombinant plasmid pgCOL 20 was grown in *Escherichia coli* strain HB 101. The chicken DNA insert was isolated from the plasmid by digestion with *Hin*dIII followed by electrophoresis in a 1% agarose gel. The 3.8×10^3 and 6.7×10^3 basepairs fragments were recovered by electrophoretic elution.

DNA restriction fragments were obtained using standard procedures and separated on polyacrylamide gels as in [10].

The nucleotide sequences of a cloned fragment of the chick proα2 collagen gene and of cloned cDNA molecules containing sequences coding for the C1 and C2 carboxyl propeptides, were determined as in [12]. Sequence data from various experiments were compared, assembled into a single sequence and examined using the computer programs in [13] which were revised for use on the Data General Nova 3 computer.

3. Results and discussion

To isolate the mannose-rich glycopeptides from the carboxyl propeptides of proα1(I) and proα2 chains, we separated the radioactively-labeled carboxyl propeptides on CM-cellulose. The peptides were digested with trypsin and the digests were chromatographed on con A–Sepharose (fig.1) to separate the carbohydrate-containing peptides from peptides that did not contain carbohydrate. A fraction of the radioactivity applied to the con A–Sepharose column was bound to the column and could be eluted with α-methylmannoside. When the material that bound to the con A–Sepharose column was chromatographed on Bio-Gel P6, the radioactivity was found to elute in a single peak (fig.2). The amino acid and neutral sugar compositions of these glycopeptides are shown in

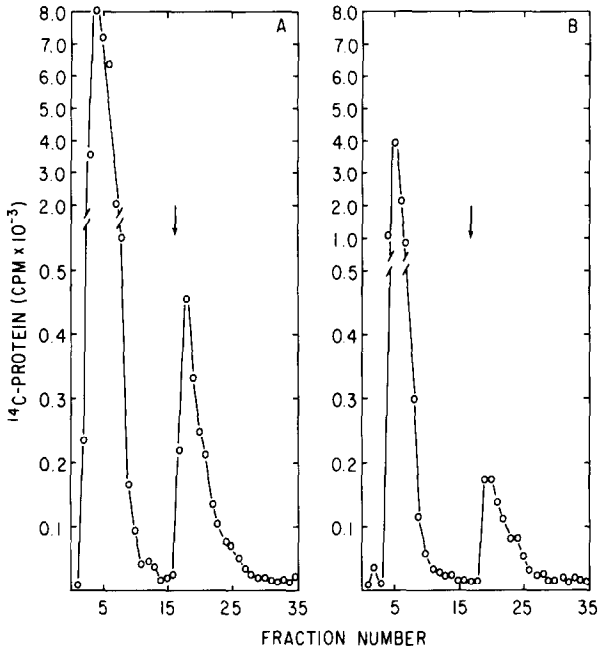

Fig.1. Isolation of the C1 (A) and C2 (B) tryptic glycopeptides by chromatography on con A–Sepharose. After the elution of peptides that did not bind to the column the glycopeptides were eluted with 0.05 M α-methylmannoside (→).

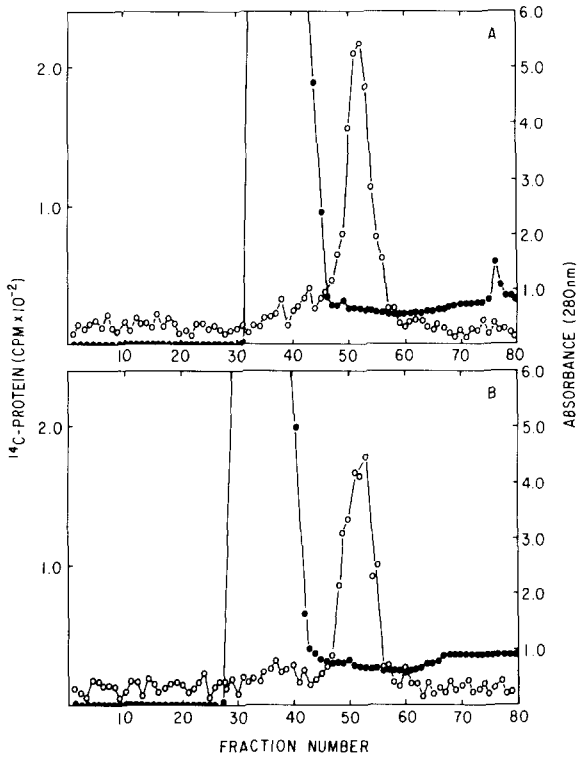

Fig.2. Chromatography of the C1 (A) and C2 (B) tryptic glycopeptides on Bio-Gel P6. The Triton X-100/SDS micelles as detected by their absorbance (•) were separated from the radioactively labeled glycopeptides (○) in this column.

Table 1
Amino acid composition of the tryptic mannose-containing glycopeptides of the C-propeptide

Amino acid	Residues/peptide (predicted from DNA sequence)		Residues/peptide (observed values) (uncorrected)	
	C1	C2	C1	C2
Asp	1	2	0.7	1.3
Thr	3	1	1.9	1.0
Ser	1	1	1.2	1.2
Glu	2	1	2.3	1.0
Pro	–	–	0.2	0.1
Gly	–	–	1.3	1.4
Ala	1	2	1.0	1.9
Cys	1	1	–	–
Val	1	–	0.9	–
Met	1	–	0.8	–
Ile	–	1	0.4	1.1
Leu	1	2	1.2	2.1
Tyr	1	1	0.8	0.9
Phe	–	–	–	–
Lys	1	1	1.0	0.9
His	1	2	0.7	1.9
Arg	–	–	0.3	–
Total	15	15	14.7	14.8
Carbohydrate				
Mannose			9	13

table 1. Both the C1- and C2-derived peptide contained mannose, with the C2-derived peptide containing more mannose residues than the C1-derived peptide. The amino-terminal sequences of the glycopeptides recovered from the P-6 column were determined by automated Edman degradation. Although the amounts of material that were available did not allow a complete sequence determination, enough peptide was obtained to establish a partial sequence for both peptides (table 2). To obtain the complete amino acid sequence of the mannose-containing region of the C1 propeptide, we have determined the nucleotide sequence of the cDNA insert in a pre-constructed recombinant plasmid containing sequences coding for the carboxyl-end of proα1(I) chains [7,10]. To determine the amino acid sequence of the carbohydrate-containing region of the C2 propeptide, we first sequenced the cDNA insert of the recombinant plasmid, pCOL 1 [8]. pCOL 1 contains sequences coding for the C2 propeptide and a portion of these sequences is shown in fig.3. For comparison, the homologous sequence of the C1-coding DNA is also shown. The analysis of the pCOL 1 sequence showed that this cDNA insert coded for only a small part of the C2 propeptide, and moreover it did not contain a carbo-

Table 2
Edman degradation of tryptic mannose-containing glycopeptides of the C-propeptide

Cycle no.	Residues predicted from DNA sequence		Residues observed	
	C1	C2	C1	C2
1	Leu	Leu	Leu	Leu
2	Met	Leu	Met	Leu
3	Ser	Ala	–	Ala
4	Thr	Asn	–	–
5	Glu	His	Glu	–
6	Ala	Ala	Ala	Ala
7	Thr	Ser	–	–
8	Gln	Gln	Gln	–

```
C1:      LEU  ARG  LEU  MET  SER  THR  GLU  ALA  THR  GLN
pCOL3:   CTG  CGC  CTG  ATG  TCC  ACC  GAG  GCC  ACC  CAG

pgCOL20: ATG  CGT  CTG  CTG  GCC  AAC  CAT  GCC  TCC  CAG
pCOL1:   ---  ---  ---  ---  ---  ---  ---  ---  ---  ---
C2:      MET  ARG  LEU  LEU  ALA  ASN  HIS  ALA  SER  GLN

C1:      ASN  VAL  THR  TYR  HIS  CYS  LYS  ASN  SER  VAL
pCOL3:   AAC  GTC  ACC  TAC  CAC  TGC  AAG  AAC  AGC  GTC

pgCOL20: AAC  ATC  ACC  TAC  CAC  TGC  AAG  AAC  AGC  ATT
pCOL1:   ---  ---  ---  ---  ---  ---  ---  ---  ---  ---
C2:      ASN  ILE  THR  TYR  HIS  CYS  LYS  ASN  SER  ILE

C1:      ALA  TYR  MET  ASP  HIS  ASP  THR  GLY  ASN  LEU
pCOL3:   GCC  TAC  ATG  GAC  CAC  GAC  ACC  GGC  AAC  CTG

pgCOL20: GCC  TAC  ATG  GAT  GAG  GAG  ACT  GGA  AAC  CTT
pCOL1:   ---  ---  ---  ---  ---  ---  ---  ---  ---  ---
C2:      ALA  TYR  MET  ASP  ASP  ASP  THR  GLY  ASN  LEU

C1:      LYS  LYS  ALA  LEU  LEU  LEU  GLN  GLY  ALA  ASN
pCOL3:   AAG  AAG  GCT  CTG  CTG  CTC  CAG  GGA  GCC  AAC

pgCOL20: AAA  AAG  GCT  GTT  ATA  CTC  CAG  GGA  TCC  AAT
pCOL1:   ---  ---  ---  ---  ---  ---  ---  ---  ---  AAT
C2:      LYS  LYS  ALA  VAL  ILE  LEU  GLN  GLY  SER  ASN
                                          ↑

C1:      GLU  ILE  GLU  ILE  ARG  ALA  GLU  GLY  ASN  SER
pCOL3:   GAG  ATC  GAG  ATC  AGG  GCC  GAA  GGA  AAC  AGC

pgCOL20: GAT  GTT  GAA  CTA  CGA  GCT  GAA  GGC  AAC  AGC
pCOL1:   GAT  GTT  GAA  CTA  CGA  GCT  GAA  GGC  AAC  AGC
C2:      ASP  VAL  GLU  LEU  ARG  ALA  GLU  GLY  ASN  SER
```

Fig.3. Nucleotide sequences from the coding strands of the proα2-coding genomic subclone pgCOL20 and the proα2-coding cDNA clone pCOL1. For comparison the homologous sequence from the proα1-coding cDNA clone pCOL3 [7] is also shown. The proα2 cDNA ends at a *Bam*H1 sites as indicated (→). The cDNA is ~180 basepairs long and although its complete nucleotide sequence has been determined (L. A. D. unpublished) only a portion of the sequence is shown here. The carbohydrate attachment sites in the C1 and C2 propeptides are indicated by the boxed-in region. The amino acid sequences derived from the nucleotide sequences are shown in the top (C1) and bottom (C2) lines.

hydrate attachment site of the structure Asn–X–Thr/Ser. We therefore, turned to a genomic proα2 subclone, pgCOL 20 (E. Avvedimento, personal communication) to obtain additional C2 sequences. In fig.3 the results obtained by sequencing a fragment of pgCOL 20 are also shown. The fragment contains sequences that are homologous with the C1 propeptide sequence and clearly overlaps with the C2-coding cDNA sequence.

The amino acid sequences obtained from the tryptic glycopeptides, together with amino acid compositions and nucleotide sequences of the genomic DNA and cDNA made it possible to verify the location of the carbohydrate attachment sites within the carboxyl propeptides of type I procollagen. With the exception of the values for glycine and cysteine, the amino acid compositions of the C1 and C2 glycopeptides were in good agreement with the compositions predicted from the nucleotide sequences (table 1). We do not consider the differences in the glycine and the cysteine values to be significant. It is likely that the presence of glycine in the analyses indicated a low level of contamination in the glycopeptides and the absence of carboxymethylated cysteine in the analyses was probably due to technical difficulties. The amounts of mannose found in the C1 and C2 propeptides [3] and the amounts of this sugar found in the isolated tryptic fragments clearly show that the C1 and C2 propeptides contain one oligosaccharide sidechain. This sidechain is located in the carboxy-terminal half of the carboxyl propeptides as shown in fig.4.

The proα2 propeptide contains ~30% more mannose than the proα1(I) propeptide. This has also been observed in labeling experiments [6]. The significance of this 'over-glycosylation' of proα2 relative to proα1(I) is unknown.

The findings that the mannose-rich oligosaccharide units in the carboxyl propeptides of type I procollagen are within regions containing the sequences –Asn–Val–Thr– and –Asn–Ile–Thr– are in complete agreement with studies indicating that the sequence –Asn–X–Ser/Thr– is the predominant structural requirement for glycosylation of asparaginyl residues by oligosaccharide transferase [14]. The functional importance of the carbohydrate in the carboxyl end of procollagen is unknown, but the carbohydrate units may provide recognition mechanisms for secretion and/or fibrillogenesis.

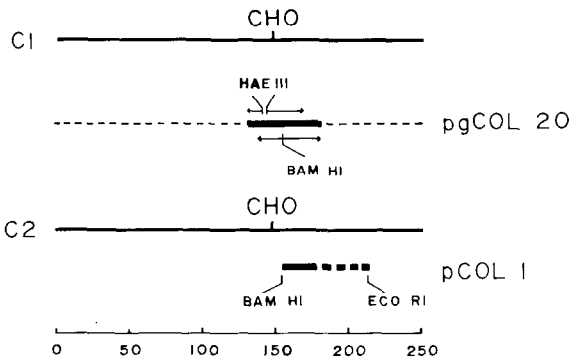

Fig.4. Diagram showing the position of the carbohydrate sidechains (CHO) in the C1 and C2 propeptides and the localization of the nucleotide sequences reported here. (→) Indicates the directions in which nucleotide sequences were determined from *Hae*III and *Bam*H1 restriction endonuclease sites. The scale at the bottom of the diagram indicates the number of amino acid residues in the carboxyl propeptides as counted from their amino-terminal ends.

Acknowledgements

We are indebted to Thomas Troy, Michael Bernard and Janey Parsons for expert technical assistance and to Dr F. Ramirez for help and advice. The study was supported in part by research grants AM 21471 and RR 09085 from the National Institutes of Health of the United States Public Health Service, grant 21–79 from the Foundation of the College of Medicine and Dentistry of New Jersey and grant PCM 7824210 from the National Science Foundation.

References

[1] Clark, C. C. and Kefalides, N. A. (1976) Proc. Natl. Acad. Sci. USA 73, 34–38.
[2] Duksin, D. and Bornstein, P. (1977) J. Biol. Chem. 252, 955–962.
[3] Olsen, B. R., Guzman, N. A., Engel, J., Condit, C. and Aase, S. (1977) Biochemistry 16, 3030–3036.
[4] Clark, C. C. and Kefalides, N. A. (1978) J. Biol. Chem. 253, 47–51.
[5] Anttinen, H., Oikarinen, A., Ryhänen, L. and Kivirikko, K. I. (1978) FEBS Lett. 87, 222–226.
[6] Clark, C. C. (1979) J. Biol. Chem. 254, 10798–10802.
[7] Showalter, A. M., Pesciotta, D. M., Eikenberry, E. F., Yamamoto, T., Pastan, I., De Crombrugghe, B., Fietzek, P. P. and Olsen, B. R. (1980) FEBS Lett. 111, 61–65.
[8] Sobel, M. E., Yamamoto, T., Adams, S. L., DiLauro, R., Avvedimento, E. V., De Crombrugghe, B. and Pastan, I. (1978) Proc. Natl. Acad. Sci. USA 75, 5846–5850.
[9] Vogeli, G., Avvedimento, E. V., Sullivan, M., Maizel, J. V. jr, Lozano, G., Adams, S. L., Pastan, I. and De Crombrugghe, B. (1980) Nucleic Acids Res. 8, 1823–1837.
[10] Yamamoto, T., Sobel, M. E., Adams, S. L., Avvedimento, V. E., DiLauro, R., Pastan, I., De Crombrugghe, B., Showalter, A., Pesciotta, D., Fietzek, P. and Olsen, B. (1980) J. Biol. Chem. 255, 2612–2615.
[11] Edman, P. and Henschen, A. (1975) in: Protein Sequence Determination (Needleman, S. B. ed) Mol. Biol. Biochem. Biophys. 8, 232–279.
[12] Maxam, A. M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560–564.
[13] Staden, R. (1979) Nucleic Acids Res. 6, 2601–2610.
[14] Hart, G. W., Brew, K., Grant, G. A., Bradshaw, R. A. and Lennarz, W. J. (1979) J. Biol. Chem. 254, 9747–9753.